Cerebral Cortex Advance Access published August 12, 2016

*Cerebral Cortex*, 2016; 1–17

doi: 10.1093/cercor/bhw240
Original Article

ORIGINAL ARTICLE

# Predicting Neural Activity Patterns Associated with Sentences Using a Neurobiologically Motivated Model of Semantic Representation

Andrew James Anderson[1], Jeffrey R. Binder[2], Leonardo Fernandino[2], Colin J. Humphries[2], Lisa L. Conant[2], Mario Aguilar[3], Xixi Wang[1], Donias Doko[1], and Rajeev D. S. Raizada[1]

[1]Brain and Cognitive Sciences, University of Rochester, NY 14627, USA, [2]Medical College of Wisconsin, Department of Neurology, Milwaukee, WI 53226, USA, and [3]Teledyne Scientific Company, Durham, NC 27703, USA

Address correspondence to Andrew James Anderson, Brain and Cognitive Sciences, University of Rochester, NY 14627, USA. Email: andrewanderson@mail.bcs.rochester.edu

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

## Abstract

We introduce an approach that predicts neural representations of word meanings contained in sentences then superposes these to predict neural representations of new sentences. A neurobiological semantic model based on sensory, motor, social, emotional, and cognitive attributes was used as a foundation to define semantic content. Previous studies have predominantly predicted neural patterns for isolated words, using models that lack neurobiological interpretation. Fourteen participants read 240 sentences describing everyday situations while undergoing fMRI. To connect sentence-level fMRI activation patterns to the word-level semantic model, we devised methods to decompose the fMRI data into individual words. Activation patterns associated with each attribute in the model were then estimated using multiple-regression. This enabled synthesis of activation patterns for trained and new words, which were subsequently averaged to predict new sentences. Region-of-interest analyses revealed that prediction accuracy was highest using voxels in the left temporal and inferior parietal cortex, although a broad range of regions returned statistically significant results, showing that semantic information is widely distributed across the brain. The results show how a neurobiologically motivated semantic model can decompose sentence-level fMRI data into activation features for component words, which can be recombined to predict activation patterns for new sentences.

**Key words:** concepts, embodiment, lexical semantics, multimodal model, semantic memory

## INTRODUCTION

Considerable progress has been made over recent decades in understanding how conceptual knowledge is represented in the human brain. In particular, functional neuroimaging studies have identified a widely distributed, large-scale network of sensory association, multimodal, and cognitive control systems that store and retrieve conceptual information (see Lau et al. 2008; Binder et al. 2009 for reviews). In addition, an extensive body of work has begun to unravel the organization of specific types of knowledge within this broad network (e.g., Kiefer and Pulvermüller 2012; Meteyard et al. 2012; Martin 2015; Fernandino et al. 2015b; Binder and Desai 2011). This progress has encouraged efforts to develop computational models that predict neural activation patterns as a function of the meaning

© The Author 2016. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

content being processed by the brain. In the first study of this kind, Mitchell et al. (2008) used word co-occurrence statistics derived from a large text corpus to model semantic content. In this type of approach, word meaning is represented as a vector of values that indicate how often the word co-occurs with other words. The fact that the word “alligator” often occurs near the word “swamp” in text samples, for example, is considered a part of the representation of “alligator”, and the fact that the vector of values representing “alligator” is similar to the vector of values for “crocodile” is taken as evidence that such representations capture meaningful semantic structure (Landauer and Dumais 1997). Mitchell et al. (2008) demonstrated that a computational model that learned the mappings between such text-based meaning representations and patterns of brain activity measured by fMRI could then be used to predict patterns of brain activity for test words not used in training the model, for which the meaning representations were known.

In the current study we expand on this and other (Devereux et al. 2010; Murphy et al. 2012; Pereira et al. 2013; Carlson et al. 2014; Anderson et al. 2015) pioneering work in two ways. First, we introduce a method for predicting patterns of neural activity arising from thinking about the meaning of an entire sentence rather than a single word. In natural language, words nearly always appear in the context of phrases and sentences rather than in isolation, thus it could be argued that isolated words are an unnatural target for understanding meaning representation in the brain. Although a great deal of neuroimaging and electrophysiological work has focused on characterizing the neural systems involved in phrase and sentence processing (see, for example, Kuperberg et al. 2000; Humphries et al. 2007; Lau et al. 2008; Friederici 2011; Pallier et al. 2011; Bemis and Pylkkänen 2012; Brennan and Pylkkänen 2012; Silbert et al. 2014), almost all studies using models to predict neural semantic representations have focused on isolated concrete nouns (exceptions are Chang et al. 2009, Wehbe et al. 2014; and Huth et al. 2016; we return to differences between the current study and these in Discussion). The challenge in devising such a predictive model arises from the fact that there are an infinite number of possible sentences with varying degrees of lexical and semantic overlap. A method for generalized prediction would require sentences to be analyzed into constituent components, followed by synthesis of a predicted activation pattern from the components.

A second innovation we propose is the use of a brain-based rather than a text-based model of meaning. A vector of word co-occurrence values is an abstract representation that contains no specific information about the qualities or experienced features of the concept itself. The highly abstract nature of such representations limits the degree to which they can be interpreted in terms of actual neural systems that could contribute to conceptual representation. Thus, even if a model composed of such features is found to predict brain activity, an account of how, in neural terms, such models work is not always forthcoming. One alternative explored in recent years is concept representation as a set of experiential attributes that reflect the sensory, motor, affective and other brain processes involved in concept learning (Gainotti et al. 2009, 2013; Crutch et al. 2012; Hoffman and Lambon Ralph 2013; Lynott and Connell 2013). In this approach, the semantic content of a given concept is estimated from ratings provided by human participants on the importance of a given modality of experience, which are taken as equivalent to modal representational systems in the brain, in learning or defining the concept. For example, concepts referring to things that make sounds (e.g., dog, horn, thunder, tuba) receive high ratings on an attribute representing auditory experience relative to things that do not make sound (e.g., cloud, flower, paper, tomato). Fernandino et al. (2015a) showed that a regression model trained to map a simple representation containing just 5 sensory–motor attributes (color, shape, visual motion, sound and manipulation) to fMRI word activation patterns could predict neural activation patterns for new words. These attributes, however, cover only a fraction of the neural representations of experience and, in particular, do not capture more abstract aspects of experience. Binder et al. (2016) recently proposed a much expanded experiential attribute model that includes sensory, motor, spatial, temporal, social, emotional, and cognitive dimensions to more comprehensively span the range of human experience, including experiences with events as well as objects, and abstract as well as concrete concepts. Here we apply this 65-dimensional model of word meaning for the first time as a basis for predicting neural activation patterns of single words.

As illustrated in Figure 1, our approach begins with estimation of the neural activity patterns associated with individual words, given only fMRI data obtained while subjects read these words embedded within various sentences. Multiple regression is then used to learn a mapping between the 65-dimensional semantic representation of these words and their estimated neural patterns, which allows neural activation patterns associated with trained and untrained words to be synthesized from their semantic representations. Finally, the neural activation pattern predicted for a given sentence is estimated by averaging the synthesized activity patterns for the individual words in the sentence.

Modeling sentences as an unordered combination of constituent words, known as a “bag-of-words” model, is an obvious simplification that ignores syntactic structure, thematic role assignment, and context effects on word meaning. As a first approximation of sentence meaning, however, the method has both theoretical and empirical justification. Sentence comprehension experiments using word priming effects (e.g., Swinney 1979; Tanenhaus et al. 1979; Till et al. 1988) suggest a time course in which a word‘s sense(s) is first activated approximately independent of its context such that multiple senses of homonyms such as ‘bat’ are jointly activated (‘mammal’ and ‘sports tool’ together), followed only later by sense selection when the appropriate meaning becomes specified by contextual information. Thus, at least part of the brain response during sentence comprehension likely reflects activation of context-independent semantic representations. Computational models suggest that feature-wise addition and multiplication of text-based semantic vectors for single words often provides a reasonable proxy for representation of word combinations (Mitchell and Lapata 2010). Based on these results from behavioral and computational studies, we hypothesized that super-position of word-level neural activation patterns would predict a similar composition of word representations in the brain activated in sentence reading.

To test this approach we collected and analyzed a large fMRI data set, acquired as participants read sentences. The sentence set was prescribed as part of the Knowledge Representation in Neural Systems project (see Materials and Methods, full listing in Table S1, and Glasgow et al. 2016). Participants were scanned as they read 240 sentences containing combinations of 141 nouns, 62 verbs, and 39 adjectives, presented across 8 scanning visits. The sentences involved interactions between humans, animals, and objects, and described situations involving entities, events, and locations with different affective connotations. Examples include “The clever scientist worked at the lab”, “The yellow bird flew over the field”, “The corn grew in spring”, “The feather was

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

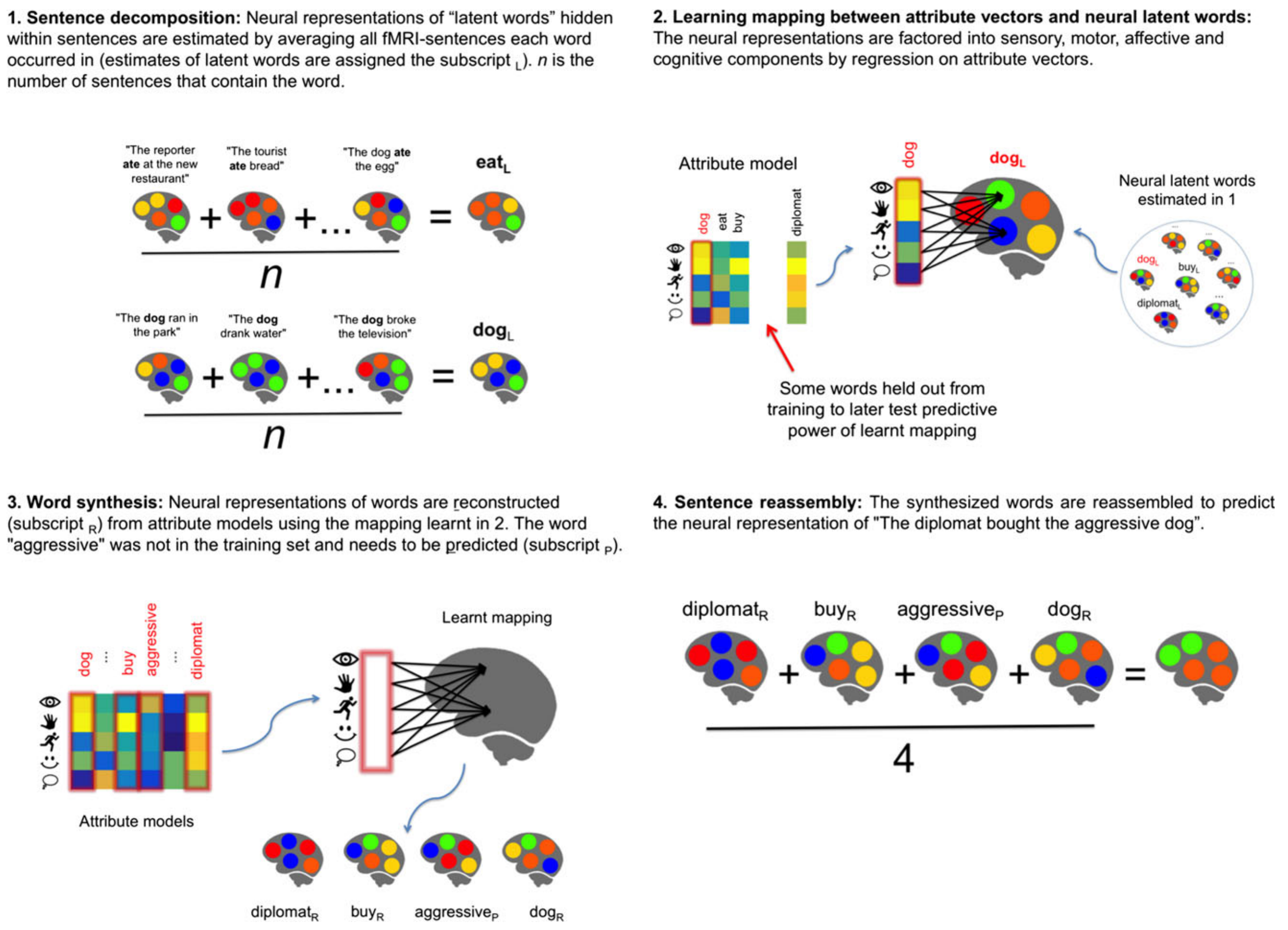


**Figure 1.** Different stages of the "sentence to word to sentence" computational approach.

blue", "The dangerous criminal stole the television", "The artist hiked on the mountain", "The cloud blocked the sun", "The banker watched the peaceful protest".

We applied our approach to predict neural activation patterns associated with compositions of trained and untrained words in untrained sentences. ROI analysis revealed that the left superior temporal sulcus (LSTS) – a structure (defined according to the Destrieux atlas in Fischl et al. 2004) that includes portions of the superior and middle temporal gyri, as well as a large portion of the angular gyrus – consistently scored highest in all our prediction tests (though significant results were also obtained within other ROIs in a cortex-level analysis). In LSTS we achieved statistically significant sentence predictions in 11/14 participants. In a series of supporting analyses we test whether discrimination accuracy degrades as words are experimentally held out from the sentence composition (e.g., a sentence of $n$ words is predicted using $n$–1 words or just a single word), and we evaluate the estimated word-level representations by testing whether they can determine which words were present in untrained sentences.

## Materials and Methods

### How do we Hypothesize Meaning is Encoded in Neural Activity?

It is now well established that different categories of visual stimuli (e.g., pictures of objects) are represented in the brain as spatially distributed neural activity patterns, with overlapping activation across categories (Haxby et al. 2001; Mitchell et al. 2008; Huth et al. 2012). It is also established that activation patterns associated with different conceptual categories (e.g., as elicited in word comprehension) partially match those elicited by viewing pictures of the word's referent (Shinkareva et al. 2011; Devereux et al. 2013; Simanova et al. 2014).

We here adopt as a working hypothesis that it is appropriate to model distributed semantic activity as a pattern of weights across a set of semantic neural features (Mitchell et al. 2008). We further hypothesize that each of these semantic neural features develops as a result of input from a specific modal neural system activated during sensory, motor, affective, and cognitive experiences. This hypothesis is broadly in line with theories of embodied or "grounded" cognition (e.g., Barsalou et al. 2008; Kiefer and Pulvermüller 2012; Meteyard et al. 2012; Binder and Desai 2011) that consider conceptual representation to involve a partial reenactment of the brain state that occurs when the concept's actual referent is experienced. For instance, embodiment theories would anticipate that, on reading the sentence "the boy kicked the ball", neural systems associated with processing the human form, lower limb biomotion, object form, and associated motion would all be activated. Obviously, semantic representations in the brain must have some degree of independence from perception/action/affective systems, otherwise all mental activity would consist of a stream of hallucinations indistinguishable from external reality.

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

Embodiment theories generally acknowledge the need for some form of "abstract" conceptual representation separate from perception and action systems (e.g., Dove 2009; Louwerse and Jeuniaux 2010; Meteyard et al. 2012; Andrews et al. 2014; Binder 2016). Moreover, such abstractions are often conceived of as linguistic representations (Glaser 1992; Barsalou et al. 2008; Dove 2009; Louwerse and Jeuniaux, 2010; Lynott and Connell 2010; Connell and Lynott 2013), providing an additional motivation for semantic models that use linguistic context as a proxy for meaning (Louwerse and Jeuniaux 2010; Andrews et al. 2014). While we do not discount the relevance of linguistic context representations (or the possibility that the brain passively accumulates linguistic distributional statistics), we propose that neural activity explained by linguistic-context models also contains substantial content associated with experiential states across many neural subsystems (Barsalou et al. 2008).

As specifically concerns sentence reading, we anticipate that early stages of comprehension will be marked by localized patterns of activity determined by weighted activation of many different neural semantic features. If the participant concentrates deeply we expect activity to be channeled approximately in proportion to feature-weight strengths to specialized experiential systems (e.g., the visual system during mental imagery of visible referents). It follows that in a case of experiential simulation, neural semantic feature weights may also describe activity visible at a macro-scale across the cortex associated with the corresponding modal systems (even though we do not attempt to model detailed simulation within these systems). This hypothesized transition from representation across many neural semantic feature weights to experiential simulation is not necessarily abrupt, and may cascade through a hierarchy of modal divergences culminating in unimodally dominant activity, consistent with Fernandino et al. (2015b). We next outline our selection of candidate neural semantic features.

## The Experiential Attribute Representation Model

The experiential attribute model (Binder et al. 2016) was developed to comprehensively span different aspects of experience as represented in neurobiological systems. Unlike other approaches to building semantic models using behavioral norming, the starting point of the attribute model is a list of well-studied modalities of neural information processing. In contrast, previous authors have had participants generate target word associates ("semantic features"), which are subsequently used directly as features in the model, or reinterpreted by the investigators in terms of experiential modalities (e.g., Cree and McRae 2003, Vinson et al. 2003).

Attributes in the model are the product of a comprehensive summary of imaging and physiological experiments (see Binder et al. 2016, 2009), and each is associated with systematic modulation in neuroimaging activity. The attributes correspond to specialized sensory/motor/affective processes; systems processing spatial, temporal, and causal information; and social cognition and abstract cognitive operations, all of which we hypothesize are activated to varying degrees in experiencing concrete and abstract entities and events. The complete list of 65 attributes is in Table 1. We model each word as a 65-dimensional feature vector that captures the strength of association between each neural attribute and that word. Target concept words were linked to attributes by having naïve participants rate the importance of each attribute for a given lexical concept.

**Table 1** List of attributes first arranged by modality, and then subdivided into individual attributes

| Dominant modality | Attribute |
|---|---|
| Vision | vision, bright, dark, color, pattern, large, small, motion, biomotion, fast, slow, shape, complexity, face, body. |
| Auditory | audition, loud, low, high, sound, music, speech. |
| Somatosensory | touch, temperature, texture, weight, pain. |
| Gustatory +Smell | taste, smell. |
| Motor | head, upper limb, lower limb, practice. |
| Attention | attention, arousal. |
| Event | duration, long, short, caused, consequential, social, time. |
| Evaluation | benefit, harm, pleasant, unpleasant. |
| Cognition | human, communication, self, cognition, number. |
| Emotion | happy, sad, angry, disgusted, fearful, surprised. |
| Drive | drive, needs. |
| Spatial | landmark, path, scene, near, toward, away. |

## Data Collection for Attribute Vectors

As per Binder et al. (2016) attribute ratings were collected on Amazon Mechanical Turk for each of the 242 content words in the set of experimental sentences. Workers were asked to rate, on a scale of 0–6, the degree to which a given lexical concept was associated with a particular type of experience (e.g., "To what degree do you think of a *football* as having a characteristic or defining color?"). The exact wording of these queries was tailored to the attribute in question and the grammatical class of the word (see Binder et al. 2016 for details). A total of 7237 rating sessions were conducted, with approximately 30 complete ratings sets (all attributes for a given word) collected for each word. Mean ratings for each word were computed, and outliers were removed by rejecting worker responses that had a Pearson's correlation coefficient of <0.5 against the mean for that particular concept (intraclass correlation). For our analyses, ratings per attribute were transformed to z-scores, and each word-vector was normed to unit length.

## Materials

All sentences were pre-selected as experimental materials for the Knowledge Representation in Neural Systems (KRNS) project (Glasgow et al. 2016, www.iarpa.gov/index.php/research-programs/krns), sponsored by the Intelligence Advanced Research Projects Activity (IARPA). The stimuli consisted of 240 written sentences containing 3–9 words and 2–5 (mean+/–sd = 3.33+/–0.76) content words, formed from different combinations of 141 nouns, 62 verbs, and 39 adjectives (242 words). Sentences were in active voice and consisted of a noun phrase followed by a verb phrase in past tense, with no relative clauses. Most sentences (200/240) contained an action verb and involved interactions between humans, animals and objects, or described situations involving different entities, events, locations, and affective connotations. The remaining 40 sentences contained only a linking verb ("was"). The entire list is in Table S1. Each word occurs a mean+/–sd [range] of 3.3+/–1.7 [1 7] times throughout the entire set of sentences and co-occurs with 8.1+/–4.3 [1 19] other unique words. The same two words rarely co-occur in more than one sentence, and 213/242 words

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

never co-occur more than once with any other single word. Forty-two sentences contained instances of words not found in any of the other 239 sentences, and 3 of these sentences contained 2 unique words. There were thus 45 words that occurred in only one sentence, of which 29 were nouns, 7 were verbs and 9 were adjectives. These 42 sentences contained a mean +/–sd of 3.57+/–0.59 content words and form a test set for sentence predictions (these sentences are segregated in Table S6c).

## Participants

Participants were 14 healthy, native speakers of English (5 males, 9 females; mean age = 32.5, range 21–55) with no history of neurological or psychiatric disorders. All were right-handed according to the Edinburgh Handedness Inventory (Oldfield 1971). Participants received monetary compensation and gave informed consent in conformity with the protocol approved by the Medical College of Wisconsin Institutional Review Board.

## Procedure

Participants took part in either 4 or 8 scanning visits. In each visit, the entire list of sentences was presented 1.5 times, resulting in 12 presentations of each sentence over the 8 visits in 10 participants, and 6 presentations over 4 visits in 4 participants. Each visit consisted of 12 scanning runs, each run containing 30 trials (one sentence per trial) and lasting approximately 6 minutes.

The stimuli were back-projected on a screen in white Courier font on a black background. Participants viewed the screen while in the scanner through a mirror attached to the head coil. Sentences were presented word-by-word using a rapid serial visual presentation paradigm. Nouns, verbs, adjectives, and prepositions were presented for 400 ms each, followed by a 200-ms inter-stimulus interval (ISI). Articles ("the") were presented for 150 ms followed by a 50-ms ISI. Mean sentence duration was 2.8 s. Words subtended an average horizontal visual angle of approximately 2.5°. A jittered inter-trial interval, ranging from 400 to 6000 ms (mean = 3200 ms), was used to facilitate deconvolution of the BOLD signal. Participants were instructed to read the sentences and think about their overall meaning. They were told that some sentences would be followed by a probe word, and that in those trials they should respond whether the probe word was semantically related to the overall meaning of the sentence by pressing one of two response keys (10% of trials contained a probe). Participants were given practice with the task outside the scanner with a different set of sentences. Response hand was counterbalanced across scanning visits.

## MRI Parameters and Preprocessing

MRI data were acquired with a whole-body 3 T GE 750 scanner at the Center for Imaging Research of the Medical College of Wisconsin. Functional T2*-weighted echoplanar images (EPI) were collected with TR = 2000 ms, TE = 24 ms, flip angle = 77°, 41 axial slices, FOV = 192 mm, in-plane matrix = 64 × 64, slice thickness = 3 mm, resulting in 3 × 3 × 3 mm voxels. T1-weighted anatomical images were obtained using a 3D spoiled gradient-echo sequence with voxel dimensions of 1 × 1 × 1 $mm^3$. fMRI data were pre-processed using AFNI (Cox 1996). EPI volumes were corrected for slice acquisition time and head motion. Functional volumes were aligned to the T1-weighted anatomical volume, transformed into a standardized space (Talairach and Tournoux 1988), and smoothed with a 6mm FWHM Gaussian kernel. The data were analyzed using a general linear model with a duration-modulated HRF. The model included one regressor for each sentence.

After preprocessing and transformation of voxel activity to z-scores, a single sentence-level fMRI representation was created for each sentence per participant by taking the voxelwise mean of all replicates of the sentence.

## Decomposing fMRI Representations of Sentences into Latent Words

All fMRI representations used in the analyses reflect processing of entire sentences as opposed to isolated words. As our modeling approach is predicated on using words as a link to 'context-invariant' neural semantic features, we introduce a strategy to extract latent representations of words from sentences. In the remaining text we use the subscript $_{\text{fMRI}}$ (e.g., $\text{sentence}_{\text{fMRI}}$) to indicate an fMRI activity pattern as opposed to other possible representations of sentences/words.

Given sentence-level fMRI data ($\text{sentence}_{\text{fMRI}}$) recorded as participants read a set of S unique and meaningful sentences, formed from a dictionary of W content words (nouns, verbs and adjectives), with some words appearing in many sentences, and some in just one, how can the fMRI patterns of individual words forming the sentences be estimated? For each of the W words, we first identify the subset of sentences in which the word occurred. Second, we estimate the latent fMRI representation of that word by taking the voxelwise mean of the fMRI patterns for all sentences in which the word occurred (we refer to this as a *latent-word*$_{\text{fMRI}}$). The result of this is that each latent-word$_{\text{fMRI}}$ estimate is built from a composite of examples of that word, complemented by/contaminated with a set of other words appearing in the same sentence context. This is illustrated in Figure 1 (stage 1).

Thus, the latent-word$_{\text{fMRI}}$ for "cow", would contain aspects of "eating", "grass" and "field", if "the cow ate grass in the field" occurred in the set of sentences. This inclusion of contextual information per se is not unreasonable given that many semantic models in computational linguistics are based entirely on word context (Landauer and Dumais 1997; Turney and Pantel 2010), and similar models have successfully been applied to explain brain data (e.g., Mitchell et al. 2008; Carlson et al. 2014; Anderson et al. 2015). For this to be appropriate, however, the sentence decomposition approach relies on having a sufficiently large set of sentences and an adequate distribution of words amongst the sentences, such that each latent-word is estimated from a reasonable word content/context ratio. What constitutes 'sufficiently large' and a 'reasonable ratio' is an empirical question. In this study we use 240 simple sentences, formed from 242 content words (nouns, verbs and adjectives), where each latent-word$_{\text{fMRI}}$ is estimated from an average of 3.3 sentences containing contextual traces of 8.2 other words. The same pair of words infrequently appears in more than one sentence (see Materials).

As it stands, this set up is not optimal, since 42 of the sentences contain words unique to the sentence set, and for those unique words the latent-word$_{\text{fMRI}}$ estimate will be an instance of a sentence$_{\text{fMRI}}$. Three sentences contain 2 unique words (the other 39 contain 1 unique word), and the estimate for both unique words within the same sentence will be identical (the same sentence$_{\text{fMRI}}$). The semantic content of these

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

semantically "contaminated" neural activity patterns, however, can be refined through a process of regression on attribute vectors (ideally removing unwanted semantic information and noise).

## Using Attribute Vectors to Synthesize Word-level Neural Activity Patterns

We used multiple regression to learn a mapping between attributes and voxels, with a training set of attribute vectors of words as independent variables and corresponding latent-words$_{fMRI}$ as dependent variables. A separate multiple regression was trained for each voxel, and all attributes were entered at once. The beta coefficients produced by these regression analyses yield brain maps, for each attribute, of the degree of modulation in activity at each voxel by the attribute. These can be thought of as basis functions relating the attribute value for a given word with the contribution of that attribute to the neural activation pattern elicited by the word (see Fernandino et al. 2015a). This is illustrated in Figure 2, see also Figure 1 (stage 2). This allows neural activity patterns for words to be synthesized, voxel by voxel, from attribute vectors using the equation below:

$$y_v = \sum_{i=1}^{65} c_{vi} a_i(w) \quad (1)$$

where $a_i(w)$ is the ith attribute for a content word $w$ and $c_{vi}$ is a parameter corresponding to the beta coefficient of the ith attribute on the $v$th voxel learnt in regression.

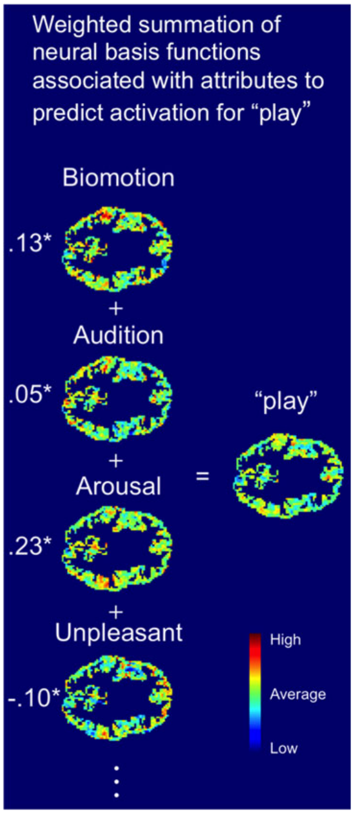


**Figure 2.** Weighted summation of brain maps (corresponding to the beta-coefficients $c_{vi}$ from Equation (1)) to synthesize neural activation for "play".

We distinguish between attribute-based syntheses of new words (that were not available in training the regression; i.e., $w$ was not in the training set) and neural activity patterns regenerated from attribute vectors for words that were trained in the regression ($w$ was in the training set), with the respective terms *predicted-words*$_{fMRI}$ and *regenerated-words*$_{fMRI}$. This difference is illustrated in Figure 1 (stage 3).

Regressing voxel activity on attributes will, in principle, factor desirable semantic content in neural activity patterns, where irrelevant semantic signal and other aspects of noise will contribute to the error. The neural activity pattern regenerated from attribute vectors in equation (2) will reconstruct the expected activity (but not the error) and ideally be 'semantically-filtered' of irrelevant semantic content and other aspects of noise. However, this improvement in regenerated-words$_{fMRI}$ over latent-words$_{fMRI}$ is expected only if the attribute vectors contain sufficiently rich semantic content to restore desirable semantic fMRI signal, which we test empirically as documented in the Results.

## Using Synthesized Word-level Activity Patterns to Estimate Sentence-level Activity

The synthesized neural activity patterns corresponding to words (generated via equation (1)) can be combined to predict compositions of words in sentences simply by averaging. We refer to this as a *predicted-sentence*$_{fMRI}$, where the voxel-level prediction for an entire sentence (incorporating equation (1)) can be formally expressed as:

$$y_v = \frac{1}{s} \sum_{j=1}^{s} \sum_{i=1}^{65} c_{vi} a_{ij}(w_j) \quad (2)$$

where $a_{ij}(w_j)$is the ith attribute for the $j$th word ($w_j$) in the predicted sentence representation (where there are S content words in the sentence), $c_{vi}$ is a parameter corresponding to the beta coefficient of the ith attribute on the $v$th voxel learnt in regression. Averaging of synthesized word activity patterns (regenerated-words$_{fMRI}$) to predict sentences is illustrated in Figure 1 (stage 4). This compositional strategy is appropriate to capture word activation that is independent of context in sentence comprehension (as is suggested to occur in the early stages of word comprehension (Till et al. 1988)). In Figure S2 we compare averaging words to estimate sentences with the obvious (and poorer performing) alternative of multiplication.

## Analysis Overview

Five analyses were undertaken. The first is a necessary supporting analysis to demonstrate that both latent-words$_{fMRI}$ (Fig. 1, stage 1) decomposed from sentences$_{fMRI}$ and word-level attribute vectors have a similar semantic structure. This verifies that it is reasonable to consider both as an operable proxy for context-invariant word representations in the brain.

The second analysis factors the decomposed latent-words$_{fMRI}$ using the attribute vectors. Focus here is placed on refining the latent-words$_{fMRI}$ to remove irrelevant semantic signal left over from the sentence decomposition process, as well as signal associated with word-form properties (e.g., orthographic and phonological features) and other noise. We verify that regression on the attributes can filter semantic features associated with words from the fMRI signal. We test this by comparing how well latent-words$_{fMRI}$ and attribute-model-regenerated-words$_{fMRI}$ are able to "spot" words in unseen fMRI

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

sentences. Importantly, this also demonstrates that we can use word-level semantic models to probe the word content of sentence-level fMRI activity.

The third analysis is the key test of our central hypothesis. It aims to demonstrate that synthesized fMRI word activation patterns can be combined to predict sentence-level fMRI activation patterns. We test where in the brain representations consistent with superpositions of multiple words can be detected by observing where sentence decomposition degrades when words are experimentally omitted from the multi-word combination.

We include a fourth analysis in Supplementary Materials. This was motivated by theories that consider the anterior temporal lobe to play a central role in conceptual representation (e.g., Patterson et al. 2007) and furthermore by theories that it is central to conceptual combination (Bemis and Pylkkänen 2012; Brennan and Pylkkänen 2012; Westerlund and Pylkkänen 2014; Zhang and Pylkkänen 2015). This analysis repeats the third analysis, specifically focusing on anterior, middle, and posterior subregions of the temporal lobe. In addition it compares averaging and multiplication as methods for building "bag-of-words" sentence representations. This comparison was motivated by previous work that has observed performance benefits to using multiplication to model word pairs using text-based computational models (Chang et al. 2009; Mitchell and Lapata 2010).

A fifth analysis examines the factorization of neural activity patterns into attributes, by identifying which attributes cumulatively received the highest regression weights (in stage #2 of Fig. 1) and how this varies across different brain regions within and across different participants' brains.

## Results

### Testing on the Whole Cortex and on Localized Brain Regions

In all analyses, activation was predicted for every voxel in the cortex and then these predictions were evaluated globally on all voxels (Cortex-level), and also within regions of interest (ROI) segmented using the Destrieux Atlas (Fischl et al. 2004). The left superior temporal sulcus (LSTS), which in the Destrieux atlas includes portions of the superior and middle temporal gyri, anterior temporal lobe, and angular gyrus, was observed posthoc to yield the strongest results across all tests, and the only ROI to reflect a statistically significant degradation in decoding accuracy caused by experimentally omitting words from the sentence model. In light of this, in the main article we list and evaluate results in detail both at cortex-level (no voxel selection within the cortex) and for LSTS. Results for all other ROIs are summarized diagrammatically with full listings in Supplementary Materials.

### Do Attribute Vectors and latent-words$_{fMRI}$ Faithfully Represent Word-level Meaning?

The first analysis was used to establish that attribute vectors are faithful representations of word meaning and also that the process of decomposing sentences into latent-words$_{fMRI}$ reliably captures semantic regularities associated with the target words. To verify that the set of latent-words$_{fMRI}$ extracted from the sentences$_{fMRI}$ captures neural activity specific to the intended word, and to jointly confirm that the attribute vectors contain discriminable word-level semantic content, we matched latent-words$_{fMRI}$ to attribute vectors using a representational similarity analysis (Anderson et al. 2016) analogous to the observed/predicted word-pair decoding algorithm that has conventionally been used to assess predictions of word-level neural activity (e.g., Mitchell et al. 2008; Chang et al. 2010; Sudre et al. 2012).

Mitchell et al.'s (2008) conventional test selects two words at a time, predicts brain-activity for these words, and then correlates predicted activity with observed activity (to give four correlation values). If the sum of correlations between the congruent predicted/observed pair exceeds the sum for the incongruent pair, decoding is a success, otherwise a failure. This process is repeated for all possible word pairs, with the mean accuracy giving a metric of success. As we have no ground truth for neural activity patterns for latent-words$_{fMRI}$ (because they are hidden in sentence-level data), this predicted/observed matching approach cannot be applied. Nevertheless, decoding can still be achieved by abstracting the matching process to representational similarity space.

Within both brain and model space, words can be re-represented in terms of their similarities to other words by correlating all word pairs within their native (brain or model) spaces. This produces two square correlation matrices of word pair similarities, one for models and the other for latent-words$_{fMRI}$. We use Pearson's correlation to generate the matrices. In model/brain similarity spaces, each word is now represented as a similarity vector of correlations with all other words, thereby allowing model and brain representations of words to be directly compared. In decoding, two test words are selected and their respective similarity vectors are drawn from both attribute model and latent-word$_{fMRI}$ correlation matrices. Entries in the similarity vectors corresponding to self-correlations (on the correlation matrix diagonal) and correlations between the two test words are removed from both test word similarity vectors to eradicate information that could give away the answer to decoding (i.e., 2 entries are removed from each vector). Decoding can then be achieved following Mitchell et al.'s conventional strategy by comparing summed correlations of congruent and incongruent matches between model and latent-word$_{fMRI}$ similarity vectors.

Rather than comparing predicted neural words to actual neural words (which we did not directly measure, as all words were embedded in sentences during scanning), this test compares word-level correlational structure within the set of attribute vectors to that within the set of latent-words$_{fMRI}$. Both attribute vectors and latent-words$_{fMRI}$ are independently constructed to capture the same thing (word-level semantic representations in the brain). If structure within model and brain data sets match each other, then we have confidence that this commonality is rooted in word-level semantic structure.

Three sentences that each had two words that were not found in any other experimental sentences were excluded from the analysis because the latent-word$_{fMRI}$ estimates for each of the respective unique words per sentence would be identical (a consequence of the sentence decomposition algorithm that we deal with in the next section). Removal of these sentences did not create any other sentences containing two unique words. Analysis was therefore on 236 words (138 nouns, 60 verbs and 38 adjectives).

Per-participant results and mean decoding accuracies in ROIs are illustrated in Figure 3. Mean+/−sd decoding accuracies for all 236 words at cortex-level were 0.71+/−0.05 (chance-level accuracy is 0.5). All participants' results were statistically significant ($p<0.05$) as determined empirically

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

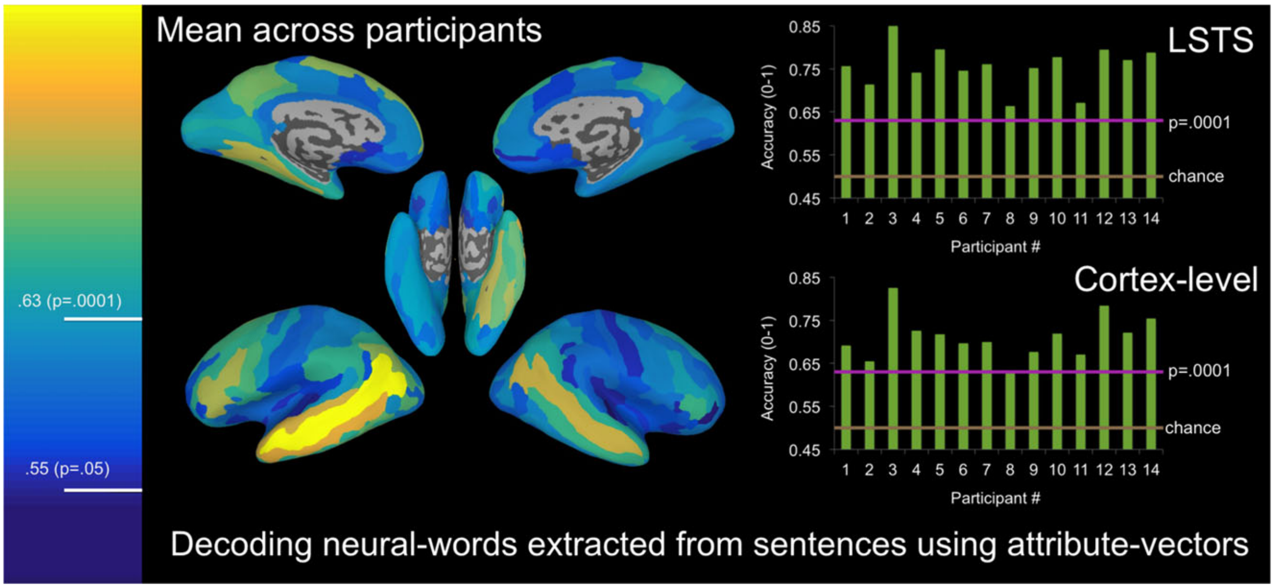


**Figure 3.** Similarity-based decoding of latent-words$_{fMRI}$ using attribute vectors. Mean decoding accuracies across participants per ROI are color-coded on a generic inflated brain surface. Critical decoding accuracies indicated in the colorbar correspond to individual-level analysis. Decoding accuracies per participant for LSTS and the cortex-level analyses are in the plots to the right (LSTS is the bright yellow strip on the lower left brain map). For participants #1–10 there were 12 replicate scans of each sentence, for #11–14 there were 6 replicates.

through permutation testing (see *Supplementary Materials*). The binomial probability of achieving 14 significant decoding results at a threshold of $p = 0.05$ is <0.0001. Comparative accuracies for LSTS were higher for each participant (mean+/−sd = 0.76+/−0.05). A comprehensive listing of decoding accuracies per ROI per participant and number of voxels per ROI is in Tables S4a/b. Regions commonly associated with semantic processing, particularly those in the left hemisphere, scored highly (e.g., middle temporal, occipito-temporal, inferior frontal, and posterior inferior parietal regions, and the precuneus).

To verify that the previous results were not driven by categorical differences in fMRI activity associated with nouns, verbs, and adjectives (there is evidence that nouns and verbs are represented in at least partially distinct systems e.g., Caramazza and Hillis 1991; Vigliocco et al. 2011), test words were partitioned into the three word class categories (N, V, and A), and the same procedure was repeated for each. Results at cortex-level and for LSTS are listed in the following format: mean+/−sd [max min] (*n*-participants significant at $p<0.05$, cumulative binomial probability of achieving $\geq n$ significant results at $p = 0.05$). Significance was empirically determined by permutation tests (*Supplementary Materials*), and critical values at $p = 0.05$ are displayed as subscripts for N, V and A. Cortex-level: $N_{0.57}$ 0.69+/−0.06, [0.81 0.57], (13/14, $p < 0.0001$); $V_{0.62}$ 0.63 +/−0.10 [0.88 0.52], (6/14, $p < 0.0001$); $A_{0.64}$ 0.64+/−0.08 [0.79 0.54]. (6/14, $p < 0.0001$). LSTS: $N_{0.57}$ 0.75+/−0.05, [0.84 0.66], (14/14, $p < 0.0001$); $V_{0.62}$ 0.72+/−0.08 [0.90 0.63], (14/14, $p < 0.0001$); $A_{0.64}$ 0.71+/−0.07 [0.79 0.55] (13/14, $p < 0.0001$). In summary, at cortex-level, noun results were significant for 13/14 participants, where verbs and adjectives were significant in approximately half of the cases. For LSTS all but one participant's test on adjectives were significant.

In closing, this section has confirmed that both attribute vectors and latent-words$_{fMRI}$ carry a significant degree of word-specific semantic information, and that word-related semantic information is available at many sites across the brain, with the strongest traces in brain regions previously associated with semantic tasks.

## Does Model-based Synthesis Improve the Semantic Specificity of Word Activation Patterns?

We anticipated that latent-words$_{fMRI}$ decomposed from sentences would be contaminated by irrelevant semantic content associated with other words (e.g., if there are only a small number of sentences referencing cows, and one of them is "the car drove past the cow", the sentence decomposition approach may exaggerate the association between cows and cars). In addition latent-words$_{fMRI}$ can be expected to contain undesirable aspects of signal associated with non-semantic word-form properties (e.g., length, letter combination statistics) and other noise. The attribute vectors should not suffer the same deficit because humans (presumably) are more reliable in producing attribute ratings that are specific to the meaning of the target word out of context (i.e., we do not expect humans typically to assign attributes associated with 'cars' to 'cows' and vice versa, or to base attribute ratings on word-form properties). We therefore proposed that model-based regeneration of latent-words$_{fMRI}$ could serve a 'semantic filtering' role by removing irrelevant semantic information and other aspects of unwanted signal associated with word-form properties and other noise. That is, since the unwanted signal is ideally only present in the fMRI data, it will not co-vary with model features and consequently will contribute to the error in regression; see *Materials and Methods*. To test this claim we compare the word-level semantic specificity of regenerated-words$_{fMRI}$ vs. latent-words$_{fMRI}$ by testing which representation is best for "spotting" words in held-out sentences$_{fMRI}$. To spot words in the held-out sentences, we used Pearson correlation to measure the similarity between the held-out-sentences$_{fMRI}$, and all fMRI word activation patterns (latent- or regenerated-words$_{fMRI}$) with the natural expectation that correlations for words that were actually present in the held-out-sentence would be highest, and the optimistic prediction that

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

correlations would be generally higher for the 'semantically-filtered' regenerated-words$_{fMRI}$.

This test was restricted to the 198 sentences formed only from words that were also present in other experimental sentences. This was to ensure that all the words in a held-out-sentence could be decomposed and regenerated from the set of other sentences. There were 197 content words (N = 112, V = 55, A = 30) in the 198 sentences. One sentence at a time was held-out for testing, and latent-words$_{fMRI}$ were estimated by decomposing the 197 training sentences. Multiple regression was used to create regenerated-words$_{fMRI}$ from the attribute vectors, and all regenerated/latent-words$_{fMRI}$ were correlated with the held-out sentence. This process was repeated until each sentence had been left out once, leaving a 198*197 matrix of sentence vs. word correlations for both latent- and regenerated-words$_{fMRI}$.

To evaluate word-spotting accuracy, for each matrix, each row of 197 sentence vs. word correlations was ranked in descending order of correlation strength and ranks were scaled between 1 and 0, where 1 is most similar and 0 is least similar. A rank score was then assigned to each sentence by looking up the ranks of the words in that sentence and taking the mean of those ranks. The mean rank score across all 198 sentences gives a composite metric of success, where the chance level percentile rank is 0.5. Statistical significance was empirically determined by permutation testing as described in *Supplementary Materials*.

Per-participant results and mean scores per ROI are illustrated in Figure 4. At cortex-level, mean rank scores for regenerated-words$_{fMRI}$ were statistically significant at $p < 0.05$ for 12/14 participants (mean+/−sd rank of 0.55+/−0.04; cumulative binomial probability $p < 0.0001$). There was however no significant difference in score between regenerated-words$_{fMRI}$ and latent-words$_{fMRI}$ (mean+/−sd = 0.55+/−0.03, t = 1.32, $p$ = 0.21, df = 13). For LSTS, mean rank scores for regenerated-words$_{fMRI}$ were higher across all participants (mean+/−sd = 0.59+/−0.04) with 13/14 participants returning significant results (cumulative binomial probability $p < 0.0001$). In this case there was a significant improvement in scores using regenerated-words$_{fMRI}$ over latent-words$_{fMRI}$ (mean+/−sd = 0.58+/−0.04; $t$ = 3.77, $p$ = 0.002, df = 13), which suggests that regeneration had indeed selectively filtered word-specific semantic signal. A comprehensive listing of rank scores per ROI per participant for regenerated/latent-words$_{fMRI}$ is in Tables S5a/b. Regional word spotting scores were highest in left hemisphere semantic regions (e.g., inferior frontal, middle temporal, and angular gyri).

To test that the word spotting result was not driven by categorical differences between nouns, verbs and adjectives, the 198*197 sentence vs. word correlation matrices were partitioned into three 198*112 (N), 157*55 (V), 96*39 (A) matrices (the number of sentence rows differs because some sentences did not contain verbs and/or adjectives), and the previous analyses were conducted separately for each matrix. Results at cortex-level and for the LSTS are presented in the following format: mean+/−sd [max min] (n-participants significant at $p < 0.05$, cumulative binomial probability of achieving $\geq n$ significant results at $p$ = 0.05). Statistical significance was empirically determined (as described in *Supplementary Materials*), and critical values at $p = 0.05$ are displayed as subscripts for N, V and A. At cortex-level, regenerated-word$_{fMRI}$ results were significant in ~40% of cases: $N_{0.53}$ 0.54+/−0.04, [0.65 0.48], (6/14, $p <$ 0.0001); $V_{0.55}$ 0.44+/−0.05 [0.69 0.49], (4/14, $p =$ 0.004); $A_{0.56}$ 0.44+/−0.04 [0.63 0.50], (6/14, $p <$ 0.0001). In LSTS results were stronger and significant in ~80% of cases: $N_{0.53}$ 0.58+/−0.04, [0.69 0.51], (13/14, $p <$ 0.0001); $V_{0.55}$ 0.61+/−0.06 [0.74 0.50], (12/14, $p <$ 0.0001); $A_{0.56}$ 0.58+/−0.04 [0.68 0.53], (8/14, $p <$ 0.0001). For nouns only, regenerated-words$_{fMRI}$ were significantly improved over latent-words$_{fMRI}$. Mean+/−sd results for latent-words$_{fMRI}$ and associated $t$-test comparisons to regenerated-words$_{fMRI}$ are as follows: $N_{0.53}$ 0.57+/−0.03, (t = 3.0, $p$ = 0.01, df = 13); $V_{0.55}$ 0.61 +/−0.06, (t = 0.66, $p$ = 0.52, df = 13); $A_{0.56}$ 0.56+/−0.05 (t = 1.42, $p$ = 0.18, df = 13).

These analyses demonstrate that it is possible to identify nouns, verbs, and adjectives in held-out fMRI sentence representations at a level significantly better than chance at both

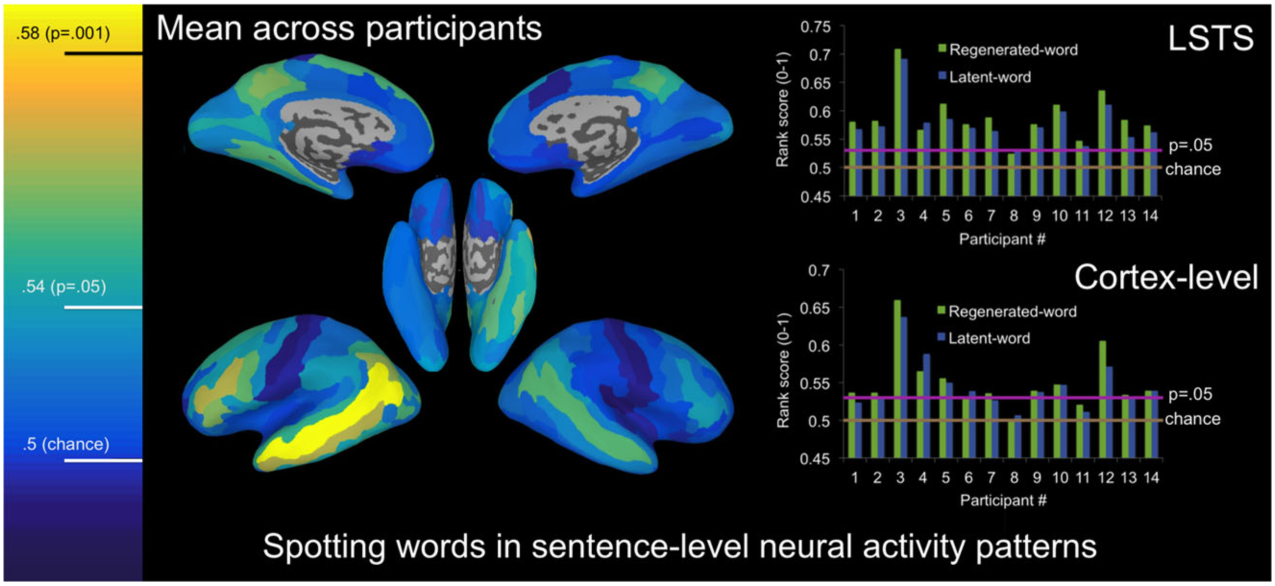


**Figure 4.** Spotting words in unseen sentences$_{fMRI}$. Mean scores across participants per ROI are color-coded on a generic inflated brain surface. Critical decoding accuracies indicated in the colorbar correspond to individual-level analysis. Scores per participant for LSTS and the cortex-level analyses are in the plots to the right (LSTS is the bright yellow strip on the lower left brain map). For participants #1–10 there were 12 replicate scans of each sentence, for #11–14 there were 6 replicates.

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

cortex-level and in localized ROIs (where results in particular in LSTS are often stronger). Furthermore, model-based regeneration of the neural activity patterns can improve the semantic information content of the fMRI signal in LSTS, but not at cortex-level. The latter result holds for nouns, but not verbs or adjectives. This could either be because the attribute vectors capture noun semantics better, or because there were less than half as many verbs and adjectives as nouns (resulting in less modeling/analytic power).

## Predicting Sentence-level fMRI Data with Word-level Attribute Vectors

Our key test is whether reconstructed neural words can be combined in a "bag-of-words" fashion (i.e., unordered and without explicit syntactic information) to predict the composition of word-level activity in sentence fMRI data. In evaluating this approach, we emphasize the following criteria:

1. Regenerated-words$_{fMRI}$ generalize to new contexts. We ensure that for all predicted-sentences$_{fMRI}$ evaluated, each word in the test sentence appears in a new context (the test sentence is unseen in training and, therefore, so is the context of each word in the sentence).
2. Predicted-words$_{fMRI}$ add semantic content to the predicted sentence representation. We test whether predicted-sentences$_{fMRI}$ formed by combining predicted-words$_{fMRI}$ with regenerated-words$_{fMRI}$ are a better match to observed-sentences$_{fMRI}$ than is a control condition where partially complete sentences were synthesized from only regenerated-words$_{fMRI}$. For example, if the held-out sentence was "The man drank coffee" and the verb "drink" was not found in any other sentences, the control partial-sentence strategy would combine only two words: regenerated-man$_{fMRI}$ and regenerated-coffee$_{fMRI}$, where the test case would combine regenerated-man$_{fMRI}$, predicted-drink$_{fMRI}$, and regenerated-coffee$_{fMRI}$. Note that this test jointly verifies that the predicted words are both viable semantic representations and viable building blocks for sentence construction. We also added in a second control test that represented sentences as the subject noun alone, with the natural prediction that the full sentences would be a better match for the neural data than the partial sentences, which in turn would be a better match than the subject nouns alone. Observing which brain regions selectively deteriorate in decoding accuracy as words are removed is evidence that the region locally represents multiple words.

To evaluate sentence prediction accuracy we used the standard predicted/observed pair matching strategy introduced by Mitchell et al. (2008). To address our test goals in unison, we focused analyses on the 42 most challenging sentences (Table S6c), which each contain either one or two words that are not present in any of the other sentences (i.e., it was necessary to synthesize either one or two predicted-words$_{fMRI}$ to complete each test sentence).

We identified all 861 unique sentence pairs from the 42 sentences and cycled through this list, leaving out two test sentences at a time. This left 238 (out of the 240) sentences to be decomposed to estimate latent-words$_{fMRI}$ and be regressed on the attribute vectors. Regenerated/predicted-words$_{fMRI}$ within the held-out sentences were synthesized, and held-out sentences were then predicted by averaging the relevant words$_{fMRI}$ (or just regenerated-words$_{fMRI}$ for the partial-sentence control). The two predicted-sentences$_{fMRI}$ were compared to each of the held-out observed-sentences$_{fMRI}$ using Pearson correlation, and the four resulting coefficients were transformed using Fisher's $r$ to $z$ transform (arctanh). If the sum of values corresponding to the correctly matched predicted/observed pairing is greater than the sum for the incorrect match, decoding is a success, otherwise a failure. Statistical significance was empirically determined by permutation testing. Two different permutation tests were run. The first randomly shuffled sentence-level vectors relative to the sentence-labels (where the label is the written sentence). The second shuffled word-level attribute vectors relative to word-labels (the written word) prior to building sentence representations (according to the original word-labels with now mismatched attribute vectors). This verified that differences in sentence length/structure were not responsible for results. Both tests are described in detail in *Supplementary Materials* and lead to the same conclusions (the same participants return significant results under both tests).

Results for each participant and mean accuracies for each ROI are illustrated in Figure 5. Results are redisplayed against scores arising from word-level permutation tests in Fig. S1. At cortex-level, sentence activations in 7/14 participants were decoded at accuracies significantly better then chance (mean +/–sd = 0.62+/–0.10; cumulative binomial probability $p < 0.0001$). Decoding was significantly weaker in the partial-sentence control case (mean+/–sd = 0.59+/–0.08; $t = 2.71$, $p = 0.02$, df = 13, 2-tail), and the subject-noun-only condition was significantly weaker than the partial-sentence (mean+/–sd = 0.56+/–0.07; $t = 2.95$, $p = 0.01$, df = 13, 2-tail).

In LSTS, performance was stronger, and sentence data from 11/14 participants were decoded at accuracies significantly better than chance (mean+/–sd = 0.70+/–0.10; cumulative binomial probability $p < 0.0001$). Accuracies were again significantly lower in the partial-sentence control condition, (mean+/–sd = 0.67 +/–0.09; $t = 5.65$, $p = 7.9e{-}5$ (df = 13, 2-tail)), and the subject-noun condition was significantly lower than the partial-sentence condition (mean+/-sd = 0.60+/–0.07; $t = 3.55$, $p = 0.004$ (df = 13, 2-tail)). Following correction for multiple comparisons (either by using false discovery rate or Bonferroni) LSTS was the only ROI to show a significant reduction in decoding accuracy between the full and partial sentence condition.

A comprehensive listing of sentence-level decoding accuracies per ROI per participant in the full-sentence and partial-sentence conditions is in Tables S6a/b. High regional decoding accuracies are apparent across a broadly similar set of ROIs as seen in the previous analyses (e.g., inferior frontal and middle temporal gyri), however the posterior cingulate gyrus also scored comparatively highly.

A fourth analysis, included in Supplementary Materials, repeated the analysis of this section on anterior, middle, and posterior subregions of LSTS and left middle temporal gyrus (LMTG). In addition, the "bag-of-words" averaging approach used to build sentence representations in this section was compared to an approach in which the voxel activation values of the constituent words were multiplied. In brief, the results indicated that there was no benefit to splitting LSTS or LMTG into subregions (indeed, statistically higher accuracy was achieved when LSTS was treated as a whole). Secondly, multiplying word activations to build sentences yielded statistically weaker performance than averaging them in all tests. This is presumably because an attribute that is important to sentence representation can be zeroed out in multiplication if even one constituent word in the sentence has a small value for that attribute.

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

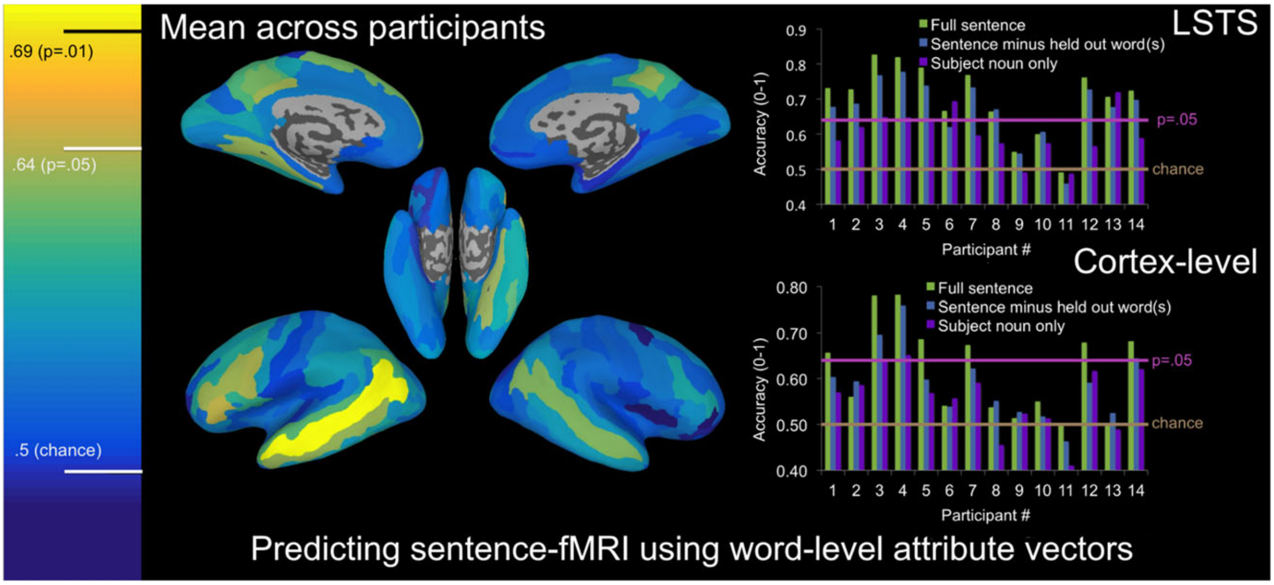


**Figure 5.** Using "bags of synthesized fMRI words" to predict and discriminate between fMRI activity elicited by unseen sentences. Mean decoding accuracies across participants per ROI are color-coded on a generic inflated brain surface. Critical decoding accuracies indicated in the colorbar correspond to individual-level analysis. Scores per participant for LSTS and the cortex-level analyses are in the plots to the right (LSTS is the bright yellow strip on the lower left brain map). For participants #1–10 there were 12 replicate scans of each sentence, for #11–14 there were 6 replicates.

These analyses demonstrate that attribute-based reconstructions of word activation patterns can be assembled to predict neural activity elicited by sentences at a level better than chance, that this approach generalizes to predict neural activity associated with novel words, and that these predictions of novel words constitute viable semantic building blocks for modeling activity elicited by sentences (because they demonstrably add semantic content to the sentence).

## Consistency in Attributes‘ Contributions Across ROIs and Participants

The list of attributes used in the model was developed to comprehensively span different aspects of experience. Given that some experiences are naturally more common than others, that different people have different experiences, that the set of situations described in the test sentences was limited, and that brain regions have different specializations, we expect differences in the profile of factor weightings within and across people's brains.

A thorough analysis of this question is beyond the scope of this work, but to get an initial impression of between-ROI and between-participant differences, we selected five ROIs that (1) supported accurate predictions in the preceding analyses; (2) are frequently implicated in conceptual tasks; and (3) are spatially spread across the network of regions commonly implicated in semantic tasks (as per Lau et al. 2008; Binder et al. 2009; Friederici 2011; Pulvermüller 2013). These were LSTS, left inferior frontal gyrus pars triangularis (LIFGtr), left angular gyrus (LAG, which in the Destrieux atlas refers to the gyral crest surrounding the posterior STS), and left posterior dorsal cingulate gyrus (LPDCing). To examine factor loadings outside these regions, an "other-cortex" ROI was built (per participant) that included all cortical voxels remaining after removing the previous ROIs bilaterally. Sentence-level prediction accuracies for the other-cortex ROI were marginally weaker than the whole-cortex-level analysis: mean+/−sd = 0.61+/−0.10 (and the same 7 participants returned significant prediction accuracies).

All 242 latent-words$_{\mathrm{fMRI}}$ were decomposed from the full set of 240 sentences$_{\mathrm{fMRI}}$, and for each participant each voxel in the cortex was regressed on the set of attribute vectors (with all attributes entered simultaneously). To build a measure of the sensitivity of the cortex or an ROI to an attribute, the squared beta coefficients from the regression for each individual attribute were summed across all relevant voxels in the cortex/ROI (in the following text these beta-coefficients are referred to as synthesis-betas):

$$sumsq_i = \sum_{v=1}^{V} c_{vi}^2 \tag{3}$$

where $sumsq_i$ is the sum of synthesis-betas for the ith attribute across all voxels $v$ in the region of interest containing a total of V voxels, and, as in equation (1), $c_{vi}$ is the synthesis-beta coefficient of the ith attribute on the $v$th voxel learnt in regression. Synthesis-beta maps for four attributes are illustrated in Figure 2.

For visualization we display positive synthesis-betas (rather than *sumsq*) because visualization of squared synthesis-betas drowns out patterns of attributes with lower loading. Cortex-level mean positive synthesis-beta profiles for all participants are overlaid in Figure 6a, where it is clear that despite variability among individuals, similar attributes tend to receive high values across participants (profiles for sum-negative or sum-squared synthesis-betas are similar and are in Fig. S3). Spearman correlation of the mean squared synthesis-beta profile between all unique pairs of participants (91 pairs) was mean+/−sd 0.83+/−0.05. All correlations were highly significant (all $p < 0.0001$). Mean squared synthesis-betas for each attribute were averaged across participants and ranked, with the highest ranking weights in descending order being: *speech*, *face*, *body*, *biomotion*, *motion*, *audition*, *pleasant*, *unpleasant*, *human*, *fast*, *happy*, *pattern*, *arousal*, *consequential*, *shape*. There could be various

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

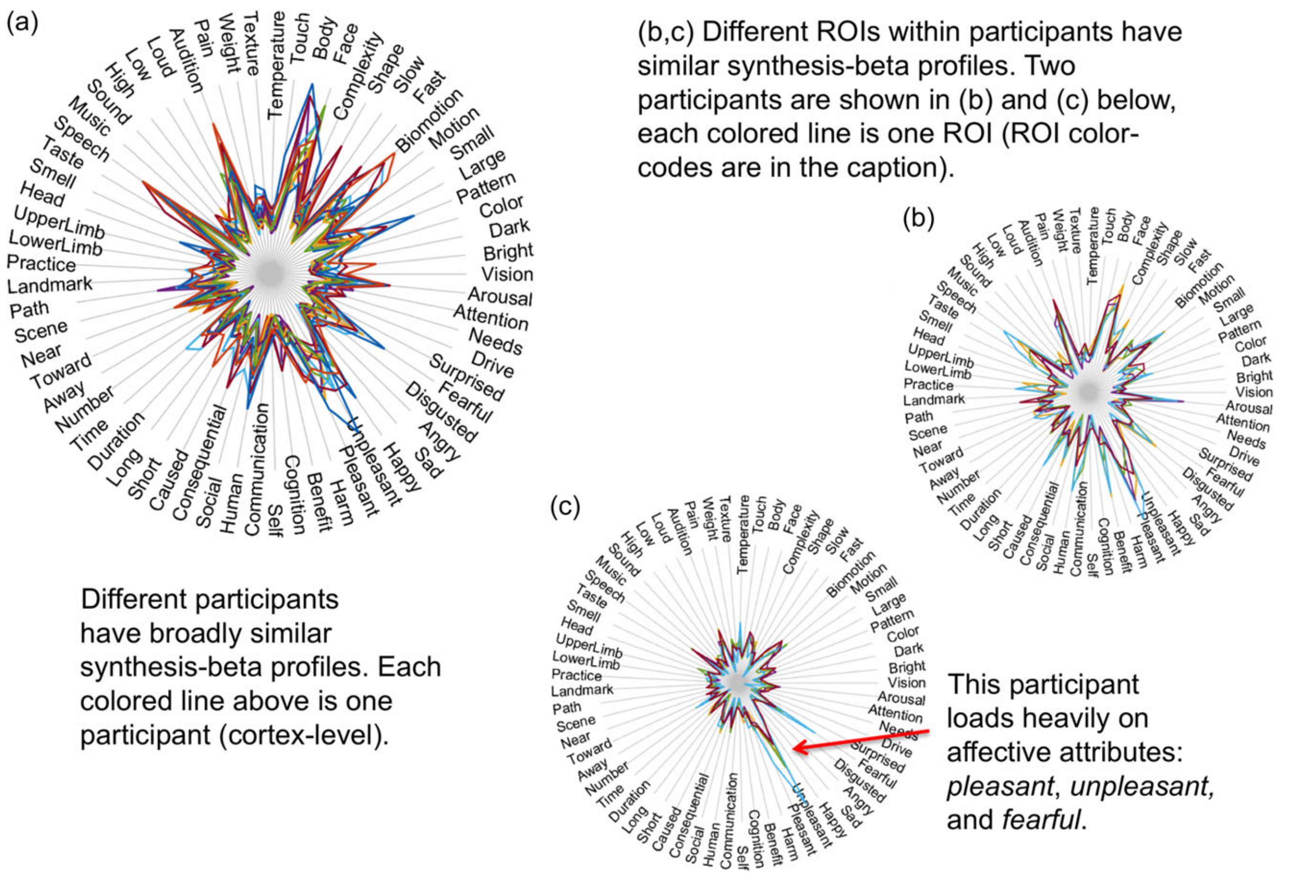


**Figure 6.** Radar plots of the sum of positive weights across voxels associated with attributes. (*a*) Cortex-level results for all participants, each participant is color-coded, colors are assigned arbitrarily to participants. (*b*,*c*). Two individuals' profiles with different ROIs' results overlaid. ROIs color-codes are LSTS (Yellow); LIFGtr (Purple); LAG (Green); LDPCing (Cyan); other-cortex (Red).

reasons that some factors were underrepresented, including that the attributes were not relevant to the sentence set, their connection to concepts cannot be reliably estimated behaviorally, or that they do not map to brain activity elicited in sentence comprehension.

Mean positive synthesis-betas in different ROIs are overlaid for the two participants with highest sentence prediction accuracies in Figure 6b,c. Within-participant profiles are similar across ROIs and also to the other-cortex ROI. A potentially interesting qualitative observation (from Fig. 6b,c) is that one participant loads comparatively heavily on affective attributes (*unpleasant, pleasant, fearful*), suggesting they had a more emotive interpretation of the test sentences. Mean squared synthesis-beta correlations between ROIs (within participants), averaged across participants, are in Table 2. Correlations are all high, and LSTS shows the strongest correlation with the other-cortex ROI (and therefore has the closest match to macro-scale neural activity patterns). Spearman correlation of mean squared synthesis-beta profiles across pairs of participants (91 unique pairs) were (mean +/–sd): LSTS 0.71+/–0.08; LIFGtr 0.68+/–0.08; LAG 0.66+/–0.12; LDPCing 0.51+/–0.12; other-cortex: 0.83+/–0.05; all $p < 0.03$ (and most $p < 0.0001$). Another qualitative observation is that the mean squared synthesis-beta profile tends to be more similar between ROIs within the same brain than across brains. One interpretation of this could be that there is a ubiquitous semantic code that is locally available across a number of brain locations and this code varies slightly more from person to person than it does between different regions of the same individual's brain.

**Table 2** Mean-squared synthesis-beta profile correlations between ROIs within participants, averaged across participants

| | LSTS | LIFGtr | LAG | LPDCing | Other-cortex |
|---|---|---|---|---|---|
| LSTS | 1.00 | 0.84 | 0.85 | 0.77 | 0.90 |
| LIFGtr | 0.84 | 1.00 | 0.78 | 0.67 | 0.86 |
| LAG | 0.85 | 0.78 | 1.00 | 0.82 | 0.86 |
| LPDCing | 0.77 | 0.67 | 0.82 | 1.00 | 0.79 |
| other-cortex | 0.90 | 0.86 | 0.86 | 0.79 | 1.00 |

## Discussion

We have introduced an approach that predicts patterns of neural activity elicited by sentence reading. Our results demonstrate that (1) neural activation specific to different nouns, verbs, and adjectives can be extracted from a large set of sentence-elicited fMRI representations; (2) word activation can be modeled using behavioral ratings that relate word meanings to neurobiologically-based experiential attributes; (3) attribute-based predictions of word-level activations can be assembled to predict activation patterns elicited by new, untrained sentences; (4) prediction accuracy is consistently higher when evaluation is focused on LSTS as opposed to the

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

whole cortex or any other segmented brain region (although multiple sites across the brain support statistically significant predictions). In the following, we discuss how these results extend previous work on conceptual combination, the interpretation of the high decoding accuracy observed in LSTS, limitations of the current approach, and potential future directions that could improve semantic models of sentences.

## How we Extend Existing Work on Conceptual Combination in the Brain

A substantial body of work has used experimental manipulation of phrase, sentence, and narrative-level stimuli to identify brain regions modulated by semantic and syntactic content of the stimuli (reviews in Friederici 2011; Hagoort and Indefrey 2014; see also Humphries et al. 2007; Graves et al. 2010; Pallier et al. 2011; Honey et al. 2012; Brennan and Pylkkänen 2012; Silbert et al. 2014). While these experiments have identified brain networks likely to be involved in processing the meaning of multi-word stimuli, details concerning how individual concepts are represented and combined within these networks have been little explored.

A few previous studies modeled neural activity elicited by word combinations and narrative level stimuli. Chang et al. (2009) predicted neural activity elicited by reading 24 adjective-noun stimuli using text-based semantic models. Vector representations of each adjective and noun were built by counting the number of times each word co-occurred with five verbs (see, hear, smell, eat, and touch) in a large text corpus. Adjectives and nouns were combined using either addition or multiplication, and the composite vectors were used to predict neural activation in selected voxels across the cortex. Although both addition and multiplication returned statistically significant results, multiplication yielded stronger predictions. Baron and Osherson (2011) demonstrated how eight concepts such as "boy", could be modeled by adding or multiplying neural activity patterns elicited by constituent concepts (e.g., "child" + "male", "child" * "male"). Addition predicted neural activation in multiple brain regions. Of these regions, the posterior cingulate and anterior temporal lobe were also predicted by multiplication, with the anterior temporal lobe alone showing an advantage for multiplication. In the current study, modeling sentences by multiplying constituent attribute vectors yielded weaker predictions, however the results reported here (where sentences described situations involving objects, actions and locations) are not directly comparable to adjective-noun combinations.

At the narrative-level, Wehbe et al. (2014) used a text co-occurrence-based semantic model together with syntax, discourse, and word-form properties to predict brain activity patterns associated with reading supra-sentential chunks of text read from a book chapter. They used these predictors to map out how different brain localities are differentially reactive to semantics, syntax, discourse and word-form. More recently Huth et al. (2016) used regression to learn a mapping between text-based semantic vectors and voxelwise neural activation elicited by listening to stories. They reduced the mapping by principal components analysis and used the first four components as the basis for an algorithm that generated semantic maps of the cortical surface. This process demonstrated regional clustering of the components across the brain surface and a similar spatial pattern of components across participants. Neurobiological interpretation of the components is ultimately ambiguous, however, since their relationship to semantic content and thus to specific brain processes is undefined.

Our analysis differs from these studies in decoding neural activation elicited by sentences, and it extends this prior work in:

*Word and feature-level detail.* We decompose sentence-level fMRI data into words, and then words into activation components associated with experiential attributes, which can be reassembled to predict sentence activation patterns. In theory, this highly analytic approach allows the generalized prediction of a very large number of sentence activation patterns for which the semantic features of constituent words are known. More generally, the current results demonstrate that hypotheses about word- and feature-level semantic content of sentences can now be tested empirically using semantic models together with sentence-level fMRI data.

*Semantic modeling.* We apply and validate a semantic model that is both built from interpretable features and comprehensively spans many aspects of experience. This is the first model of its kind that attempts to connect word meaning with the high-dimensional complexity of experiential representation in the brain. The results provide strong initial evidence that structure encoded in the semantic model is also present in brain activation patterns.

## On Interpreting the Relationship Between the Attribute Model and Neural Activity

While our results show that semantic structure across the attributes correlates with neural semantic structure, this does not automatically entail that the neural semantic code is built out of precisely the same attributes as the model. However, given the principled design of the model, and previous results from a study using five similar attributes to analyze the neural representation of isolated words (Fernandino et al. 2015a,b), a correspondence between the attributes and the neural systems on which they are based seems likely. However, future work will be necessary to clarify the nature and extent of this correspondence (e.g., whether a high attribute score on the *lower limb* attribute for a target word like "kick" predicts lower limb related brain activity, as observed by Hauk et al., 2004). In the meantime, our results show that the attribute model provides a flexible way of predicting the components that contribute to semantic representations in the brain. This view is consistent with theories considering conceptual representations to be partly embodied in modal systems (e.g., Barsalou et al. 2008; Binder and Desai, 2011; Kiefer and Pulvermüller, 2012; Meteyard et al., 2012).

As we have not directly compared decoding performance with other semantic models (e.g., Mitchell et al. 2008; Devereux et al. 2010; Murphy et al. 2012; Huth et al. 2012; Pereira et al. 2013; Anderson et al. 2013; Bruffaerts et al. 2013; Carlson et al. 2014; Wehbe et al. 2014; Fernandino et al. 2015a; Anderson et al. 2015; Huth et al. 2016), we make no claim about the superiority or otherwise of the model to other posited models of conceptual representation for the purpose of decoding brain activity. In work in progress we are exploring comparison and combination of the attribute model with state-of-the-art text-based computational semantic models (e.g., Baroni et al. 2014). To foreshadow future results, both model types are competitive in similar tests to those reported here, and they carry complementary information (as we might expect from Andrews et al. 2009).

## Importance of the Left Superior Temporal Sulcus and Surrounding Cortex

We have presented new evidence that when sentences are read, semantic representations associated with multiple words

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

are activated in LSTS, and that this activation can be predicted using attribute ratings. For consistency with the Destrieux atlas nomenclature we use the term STS to denote this region, however it is important to note that the ROI also includes large portions of the angular gyrus, middle and superior temporal gyri, and lateral anterior temporal lobe. Furthermore, the adjacent ROIs labeled angular gyrus and middle temporal gyrus also showed high levels of decoding accuracy. Previous work has linked all of these regions with semantic and syntactic processes (e.g., Lau et al. 2008; Binder et al. 2009; Friederici 2011; Pulvermüller 2013; Humphries et al. 2007; Pallier et al. 2011; Honey et al. 2012; Silbert et al. 2014), and lesions in this region are known to disrupt sentence processing (Dronkers et al. 2004; Magnusdottir et al. 2013; Thothathiri et al. 2012). Various subregions of the LSTS have also been linked (Hein and Knight 2008; Liebenthal et al. 2014) to a diverse array of other tasks that include speech perception, theory of mind, audio-visual integration, biological motion perception, and face processing (tasks that we note bear striking similarity to the attributes scoring highly in Fig. 6). Interpreted in the light of this prior evidence, our results are consistent with a central role for the STS and surrounding cortex (anterior temporal lobe, middle temporal gyrus, and angular gyrus) in the representation of lexical and sentential conceptual content. As noted by several authors, these regions are important "convergence zones" for multisensory processing streams (Beauchamp et al. 2004; Cavada and Goldman-Rakic 1989a, 1989b; Jones and Powell 1970; Seltzer and Pandya 1994) and serve as connectivity "hubs" linking multiple distributed networks (Achard et al. 2006; Sepulcre et al. 2012). Notably, other posterior cortical regions with similar convergence zone and hub characteristics – the precuneus, posterior cingulate gyrus, and left parahippocampal gyrus – also showed relatively high levels of decoding accuracy. All of these regions were previously associated with semantic processing in a large meta-analysis of neuroimaging studies (Binder et al. 2009).

## Limitations of the Present Approach, and Ways it Could be Extended

Despite this first progress on sentence-level decoding, many issues remain to be explored in future work. For example, word order is a basic factor that determines meaning (e.g., "boat house" and "house boat" do not mean the same thing), and word combinations create meaning in complex ways that reflect the intrinsic semantic properties of each word (e.g., a "cancer therapy" is a therapy for cancer, but a "water therapy" is not a therapy for water) and context-specific idiomatic meaning (e.g., a "red car" is the color red but this is not the case for "red army" or "red herring").

In this article, we have stopped short of attempting to capture the specific senses of meaning brought by the interactions of words in context. In the longer term it is desirable to construct more sophisticated methods that can combine word-level vectors to estimate specific aspects of meaning in context and, unlike the "bag-of-words" model, capture the effects of syntax and word morphology. One potential method would be to train computational models to learn how words modify each other when they appear in context (e.g., Baroni and Zamparelli 2010; Paperno et al. 2014).

In the immediate term, and directly addressable using our existing framework, attribute vectors can be estimated at word-pair, phrase, sentence, or even supra-sentence level simply by having people rate these different targets. In contrast, it is difficult to collect ground-truth text-based semantic vectors for larger phrase and sentence-level constructs, where occurrences of specific sentences, even in huge bodies of text, are often rare or non-existent

A second benefit of the attribute rating approach is that it is comparatively simple to collect person-specific attribute ratings for words and sentences, as would be useful for predicting individual differences in neural semantic representation. In contrast, to our knowledge all computational semantic models applied to brain data have been built at group-level (e.g., built from the text written by many authors in digital data repositories). Although in principle person-specific models could be compiled based on personal document stores and photographs, it is less easy to guarantee that any specific person will have accumulated sufficient data to construct such a model.

Experiential attribute models also have limitations, however. Under ideal circumstances, the set of attributes coded in the model reflects current neurobiological knowledge of brain function, yet there is no guarantee that this set is complete, since neurobiological knowledge is likely to evolve. Because the attributes are linguistic descriptions, it may not be practical or possible to identify dimensions of experience that are highly nonverbal or inaccessible to conscious awareness. A practical disadvantage is that it is costly and time consuming to collect human ratings on a massive scale (compared to running a computational algorithm on a digital data repository) even given modern internet crowdsourcing tools that facilitate the collection of such data. Furthermore, future refinements to the model (e.g., adding features) require new ratings to be collected. It may therefore prove valuable to leverage the combined benefits of semantic models from different sources (Andrews et al. 2009; Anderson et al. 2015).

Aside from syntax and morphology, the model presented in the current article does not account for detailed aspects of experiential simulation (such as mental imagery) that may be invoked in conceptual tasks (e.g. Louwerse and Hutchinson 2012), episodic memory (e.g. Hassabis et al. 2007), inferences and theory-of-mind (Hagoort and Indefrey 2014). New grounded semantic models, such as those derived from natural image statistics, have seen early success decoding visual aspects of neural activity elicited by reading (Anderson et al. 2015), and in general provide methods to target modally grounded aspects of neural representations (e.g. see also Kiela and Clark, 2015 for audio-based representations). Incorporating pragmatic inference into models is a further challenge. Certain types of inference may be amenable to modeling using experiential attribute ratings. For instance it is reasonable to conjecture that the modulation of affect (and associated changes in brain activity) invoked by sentences such as "The grandfather kicked the baby" in contrast with "The baby kicked the grandfather" as observed by Frankland and Greene (2015) could be captured by the experiential attribute model. Other inferences that require an understanding of the mental state of others, thus enabling "it's hot in here" to be selectively recognized as an indirect request to open a window in appropriate circumstances (van Ackeren et al. 2012), however, may require additional techniques to model.

In conclusion, this article has presented an approach that decomposes fMRI sentences into words, and words into embodied neural semantic features, and then reassembles them to predict new words and sentences. This has enabled statistically significant prediction of fMRI activation patterns elicited by reading sentences across a broad range of cortical regions and in particular the LSTS. The results provide initial

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

validation for the experiential attribute model and a foundation for modeling the neural representation of sentence meaning, which has many opportunities for extension.

## Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/

## Funding

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory under grant FA8650-14-C-7357. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government. Author R.D.S. R. was also supported in part by NSF Award 1228261.

## Notes

We thank two anonymous reviewers for their insightful comments and suggestions. *Conflict of Interest:* None declared.

## References

Achard S, Salvador R, Whitcher B, Suckling J, Bullmore E. 2006. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. J Neurosci. 26:63–72.

Anderson AJ, Bruni E, Bordignon U, Poesio M, Baroni M. 2013. Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013); Seattle, WA: Association for Computational Linguistics. pp. 1960–1970.

Anderson AJ, Bruni E, Lopopolo A, Poesio M, Baroni M. 2015. Reading visually embodied meaning from the brain: visually grounded computational models decode visual-object mental imagery induced by written text. Neuroimage. 120: 309–322.

Anderson AJ, Zinzser BD, Raizada RDS. 2016. Representational similarity encoding for fMRI: pattern-based synthesis to predict brain activity using stimulus-model-similarities. NeuroImage. 128:44–53.

Andrews M, Vigliocco G, Vinson D. 2009. Integrating experiential and distributional data to learn semantic representations. Psychol Rev. 116(3):463–498.

Andrews M, Frank S, Vigliocco G. 2014. Reconciling embodied and distributional accounts of meaning in language. Top Cogn Sci. 6:359–370.

Baron SG, Osherson D. 2011. Evidence for conceptual combination in the left anterior temporal lobe. Neuroimage. 55:1847–1852. doi:10.1016/j.neuroimage.2011.01.066.

Baroni M, Zamparelli R. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2010); East Stroudsburg PA: Association for Computational Linguistics, pp. 1183–1193.

Baroni M, Dinu G, Kruszewski G. 2014. Dont count, predict! A systematic comparison of context-counting vs.context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol 1: Long Papers); Baltimore, Maryland, USA: Association for Computational Linguistics.

Barsalou LW, Santos A, Simmons WK, Wilson CD. 2008.Language and simulation in conceptual processing. In: De Vega M, Glenberg AM, Graesser AC, editor. Symbols, embodiment, and meaning. Oxford: Oxford University Press. pp. 245–283.

Beauchamp MS, Lee KE, Argall BD, Martin A. 2004. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron. 41:809–823.

Bemis DK, Pylkkänen L. 2012. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. Cereb Cortex. 23 (8):1859–1873. doi:10.1093/cercor/bhs170.

Binder JR. 2016. In defense of abstract conceptual representations. Psychonomic Bulletin & Review, 23. doi:10.3758/s13423-015-0909-1

Binder JR, Desai RH, Graves WW, Conant LL. 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex. 19:2767–2796.

Binder JR, Desai RH. 2011. The neurobiology of semantic memory. Trends Cogn Sci. 15(11):527–536.

Binder JR, Conant LL, Humphries CJ, Fernandino L, Simons S, Aguilar M, Desai R. 2016. Toward a brain-based componential semantic representation. Cogn Neuropsychol. http://dx.doi.org/10.1080/02643294.2016.1147426.

Brennan J, Pylkkänen L. 2012. The time-course and spatial distribution of brain activity associated with sentence processing. NeuroImage. 60(2):1139–1148. doi:10.1016/j.neuroimage.2012.01.030.

Bruffaerts R, Dupont P, Peeters R, De Deyne S, Storms G, Vandenberghe R. 2013. Similarity of fMRI activity patterns in left perirhinal cortex reflects similarity between words. J Neurosci. 33(47):18597–18607.

Caramazza A, Hillis A. 1991. Lexical organization of nouns and verbs in the brain. Nature. 349(6312):788–90.

Carlson TA, Simmons RA, Kriegeskorte N, Slevc LR. 2014. The emergence of semantic meaning in the ventral temporal pathway. J Cogn Neurosci. 26(1):120–131.

Cavada C, Goldman-Rakic PS. 1989a. Posterior parietal cortex in rhesus monkey: i. Parcellation of areas based on distinctive limbic and sensory corticocortical connections. J Comp Neurol. 287:393–421.

Cavada C, Goldman-Rakic PS. 1989b. Posterior parietal cortex in the rhesus monkey: ii. Evidence for segregated corticocortical networks linking sensory and limbic areas with the frontal lobe. J Comp Neurol. 287:422–445.

Chang KM, Cherkassky VL, Mitchell TM, Just MA. 2009. Quantitative modeling of the neural representation of adjective-noun phrases to account for fMRI activation. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore: Association for Computational Linguistics. pp. 638–646.

Chang KM, Mitchell TM, Just MA. 2010. Quantitative modeling of the neural representations of objects: how semantic feature norms can account for fMRI activation. Neuroimage. 56:716–727.

Connell L, Lynott D. 2013. Flexible and fast: linguistic shortcut affects both shallow and deep conceptual processing. Psychon Bull Rev. 20:542–550.

Cox RW. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res. 29:162–173.

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

Cree GS, McRae K. 2003. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). J Exp Psychol Gen. 132(2):163–201.

Crutch SJ, Williams P, Ridgway GR, Borgenicht L. 2012. The role of polarity in antonym and synonym conceptual knowledge: Evidence from stroke aphasia and multidimensional ratings of abstract words. Neuropsychologia. 50:2636–2644.

Devereux B, Kelly C, Korhonen A. 2010. Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In: Murphy B, Chang KK, Korhonen A. editors. Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics. Los Angeles, USA: Association for Computational Linguistics. pp. 70–78.

Devereux BJ, Clarke A, Marouchos A, Tyler LK. 2013. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. J Neurosci. 33(48):18906–18916.

Dove GO. 2009. Beyond perceptual symbols: a call for representational pluralism. Cognition. 110:412–431.

Dronkers NF, Wilkins DP, Van Valin RD, Redfern BB, Jaeger JJ. 2004. Lesion analysis of the brain areas involved in language comprehension. Cognition. 92:145–177.

Fernandino L, Humphries CJ, Seidenberg MS, Gross WL, Conant LL, Binder JR. 2015a. Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. Neuropsychologia. doi:10.1016/j.neuropsychologia.2015.04.009.

Fernandino L, Binder JR, Desai RH, Pendl SL, Humphries CJ, Gross WL, Conant LL, Seidenberg MS. 2015b. Concept representation reflects multimodal abstraction: a framework for embodied semantics. Cereb Cortex. doi:10.1093/cercor/bhv02.

Fischl B, van der Kouwe A, Destrieux C, Halgren E, Segonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D, et al. 2004. Automatically parcellating the human cerebral cortex. Cereb Cortex. 14:11–22.

Frankland SM, Greene JD. 2015. An architecture for encoding sentence meaning in left mid-superior temporal cortex. Proc Natl Acad Sci USA. 112(37):11732–11737. doi:10.1073/pnas.1421236112.

Friederici AD. 2011. The brain basis of language processing: from structure to function. Physiol Rev. 91(4):1357–1392.

Gainotti G, Ciaraffa F, Silveri MC, Marra C. 2009. Mental representation of normal subjects about the sources of knowledge in different semantic categories and unique entities. Neuropsychology. 23(6):803–812.

Gainotti G, Spinelli P, Scaricamazza E, Marra C. 2013. The evaluation of sources of knowledge underlying different conceptual categories. Fron Hum Neurosci. 7:40.

Glaser WR. 1992. Picture naming. Cognition. 42:61–105.

Glasgow K, Roos M, Haufler A, Chevillet M, Wolmetz M. 2016. Evaluating semantic models with word-sentence relatedness. arXiv:1603.07253.

Graves W, Binder JR, Desai R, Conant L, Seidenberg MS. 2010. Neural correlates of implicit and explicit conceptual combination. NeuroImage. 53(2):638–646. doi:10.1016/j.neuroimage.2010.06.055.

Hagoort P, Indefrey P. 2014. The neurobiology of language beyond single words. Ann Rev Neurosci. 37:347–362. doi:10.1080/17470218.2015.1038280.

Hassabis D, Kumaran D, Maguire EA. 2007. Using imagination to understand the neural basis of episodic memory. J Neurosci. 27(52):14365–14374.

Hauk O, Johnsrude I, Pulvermüller F. 2004. Somatotopic representation of action words in human motor and premotor cortex. Neuron. 41(2):301–7.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 293:2425–2430.

Hein G, Knight R. 2008. Superior temporal sulcus—it's my area: or is it?. J Cogn Neurosci. 20:2125–2136.

Hoffman P, Lambon Ralph MA. 2013. Shapes, scents and sounds: quantifying the full multi-sensory basis of conceptual knowledge. Neuropsychologia. 51:14–25.

Honey CJ, Thompson CR, Lerner Y, Hasson U. 2012. Not lost in translation: neural responses shared across languages. J Neurosci. 32(44):15277–15283. doi:10.1523/JNEUROSCI.1800-12.2012.

Humphries C, Binder JR, Medler DA, Liebenthal E. 2007. Time course of semantic processes during sentence comprehension. Neuroimage. 36:924–932. doi:10.1016/j.neuroimage.2007.03.059.

Huth AG, Nishimoto S, Vu AT, Gallant JL. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron. 76(6):1210–1224.

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature. 532:453–458.

Jones EG, Powell TSP. 1970. An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. Brain. 93:793–820.

Kiela D, Clark S. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. Proceedings of the Empirical Methods in Natural Language Processing Conference Lisbon, Portugal (EMNLP 2015); Association for Computational Linguistics. pp. 2461–2470.

Kiefer M, Pulvermüller F. 2012. Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. Cortex. 48:805–825.

Kuperberg GR, McGuire PK, Bullmore ET, Brammer MJ, Rabe-Hesketh S, Wright IC, Lythgoe DJ, Williams SC, David AS. 2000. Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: an fMRI study. J Cogn Neurosci. 12(2):321–41.

Landauer T, Dumais S. 1997. A solution to Plato‘s problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev. 104(2):211–240.

Lau EF, Phillips C, Poeppel D. 2008. A cortical network for semantics: (de)constructing the N400. Nat Rev Neurosci. 9:920–933.

Liebenthal E, Desai, RH, Humphries C, Sabri M. Desai A. 2014. The functional organization of the left STS: a large scale meta-analysis of PET and fMRI studies of healthy adults. Front. Neurosci. 8:289.

Louwerse MM, Hutchinson S. 2012. Neurological evidence linguistic processes precede perceptual simulation in conceptual processing. Front Psychol. 3:385. doi:10.3389/fpsyg.2012.00385.

Louwerse MM, Jeuniaux P. 2010. The linguistic and embodied nature of conceptual processing. Cognition. 114:96–104.

Lynott D, Connell L. 2013. Modality exclusivity norms for 400 nouns: the relationship between perceptual experience and surface word form. Behav Res. 45:516–526.

Lynott D, Connell L. 2010. Embodied conceptual combination. Front Psychol. 1:212. doi:10.3389/fpsyg.2010.00212.

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016

Magnusdottir S, Fillmore P, den Ouden DB, Hjaltason H, Rorden C, Kjartansson O, Bonilha L, Fridriksson J. 2013. Damage to left anterior temporal cortex predicts impairment of complex syntactic processing: a lesion-symptom mapping study. Hum Brain Mapp. 34(10):2715–2723.
Martin A. 2015. GRAPES—grounding representations in action, perception, and emotion systems: how object properties and categories are represented in the human brain. Psychon Bull Rev. 1–12.
Meteyard L, Cuadrado SR, Bahrami B, Vigliocco G. 2012. Coming of age: a review of embodiment and the neuroscience of semantics. Cortex. 48:788–804.
Mitchell J, Lapata M. 2010. Composition in distributional models of semantics. Cogn Sci. 34(8):1388–1439.
Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, Mason RA, Just MA. 2008. Predicting human brain activity associated with the meaning of nouns. Science. 320:1191–1195.
Murphy B, Talukdar P, Mitchell T. 2012. Selecting Corpus-Semantic Models for Neurolinguistic Decoding. In Proceedings of First Joint Conference on Lexical and Computational Semantics (*SEM). Montreal Canada. Assoc Comput Linguist. pp. 114–123.
Oldfield RC. 1971. The assessment and analysis of handedness: the Edinburgh Inventory. Neuropsychologia. 9:97–113.
Pallier C, Devauchelle A-D, Dehaene S. 2011. Cortical representation of the constituent structure of sentences. Proc Natl Acad Sci USA. 108(6):2522–2527.
Paperno D, Pham N, Baroni M. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014); East Stroudsburg PA: Association for Computational Linguistics. pp. 90–99.
Patterson K, Nestor PJ, Rogers TT. 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. Nat Rev Neurosci. 8:976–987.
Pereira F, Botvinick M, Detre G. 2013. Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. Artif Intell. 194:240–252.
Pulvermüller F. 2013. How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. Trends Cogn Sci. 17(9):458–470.
Seltzer B, Pandya DN. 1994. Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. J Comp Neurol. 343:445–463.
Sepulcre J, Sabuncu MR, Yeo TB, Liu H, Johnson KA. 2012. Stepwise connectivity of the modal cortex reveals the multimodal organization of the human brain. J Neurosci. 32:10649–10661.
Simanova I, Hagoort P, Oostenveld R, Van Gerven MAJ. 2014. Modality-independent decoding of semantic information from the human brain. Cereb Cortex. 24:426–434.
Shinkareva SV, Malave VL, Mason RA, Mitchell TM, Just MA. 2011. Commonality of neural representations of words and pictures. Neuroimage. 54:2418–2425.
Silbert LJ, Honey CJ, Simony E, Poeppel D, Hasson U. 2014. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proc Natl Acad Sci USA. 111(43):E4687–96.
Sudre G, Pomerleau D, Palatucci M, Wehbe L, Fyshe A, Salmelin R, Mitchell T. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. Neuroimage. 62:451–463.
Swinney DA. 1979. Lexical access during sentence comprehension: (Re)consideration of context effects. J Verb Learn Verb Behav. 18:645–659.
Talairach J, Tournoux P. 1988. Co-planar stereotaxic atlas of the human brain. In: 3-Dimensional proportional system: an approach to cerebral imaging. New York: Thieme p. 122.
Tanenhaus MK, Leiman JM, Seidenberg MS. 1979. Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. J Verb Lear Verb Behav. 18:427–440.
Thothathiri M, Kimberg DY, Schwartz MF. 2012. The neural basis of reversible sentence comprehension: evidence from voxel-based lesion-symptom mapping in aphasia. J Cogn Neurosci. 24:212–222.
Till RE, Mross EF, Kintsch W. 1988. Time course of priming for associate and inference words in a discourse context. Mem Cogn. 16:283–298.
Turney P, Pantel P. 2010. From frequency to meaning: vector space models of semantics. J Artific Intell Res. 37:141–188.
van Ackeren MJ, Casasanto D, Bekkering H, Hagoort P, Rueschemeyer S-A. 2012. Pragmatics in action: indirect requests engage theory of mind areas and the cortical motor network. J Cogn Neurosci. 24(11):2237–47.
Vigliocco G, Vinson DP, Druks J, Barber H, Cappa SF. 2011. Nouns and verbs in the brain: a review of behavioural, electrophysiological, neuropsychological and imaging studies. Neurosci Biobehav Rev. 35(3):407–26.
Vinson DP, Vigliocco G, Cappa S, Siri S. 2003. The breakdown of semantic knowledge: insights from a statistical model of meaning representation. Brain Lang. 86(3):347–365.
Wehbe L, Murphy B, Talukdar P, Fyshe A, Ramdas A, Mitchell T. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. Plos One. 9(11):e11257.
Westerlund M, Pylkkänen L. 2014. The role of the left anterior temporal lobe in semantic composition vs. semantic memory. Neuropsychologia. 57:59–70. doi:10.1016/j.neuropsychologia.2014.03.001.
Zhang L, Pylkkänen L. 2015. The interplay of composition and concept specificity in the left anterior temporal lobe: an MEG study. Neuroimage. 111:228–240. doi:10.1016/j.neuroimage.2015.02.028.

Downloaded from http://cercor.oxfordjournals.org/ at Serials Dept / Cornell University Medical Library on August 14, 2016